Knowledge Sources for Natural Language Processing

Robert H. Baud, PhD, Anne-Marie Rassinoux, PhD, Christian Lovis, MD, Judith Wagner,
Vincent Griesser, MD, Pierre-André Michel, Jean-Raoul Scherrer, MD
Division d'Informatique Médicale
University Hospital of Geneva, Switzerland

This paper aims at reviewing the problem of feeding Natural Language Processing (NLP) tools with convenient linguistic knowledge in the medical domain. A syntactic approach lacks the potential to solve a number of typical situations with ambiguities and is clearly insufficient for quality treatment of natural language. On the other hand, a conceptual approach relies on some modelling of the domain, of which the elaboration is a long-term process and where the ultimate solutions are far from being recognised and universally accepted. In-between is the beauty of the compromise. How can we significantly improve the coverage of linguistic knowledge in the years to come?

INTRODUCTION

NLP systems using 10'000 words with a good coverage of the concerned domain are rare. A few authors have achieved such a goal under the form of prototypes taking advantage of existing corpora of linguistic knowledge. This is the best way to proceed if we do not want to reinvent the wheel. Reuse of existing sources is then a must.

This paper will consider in turn the different sources which are available. They will be considered according to the kind of linguistic knowledge they can provide, in order to highlight their specificity.

CATEGORISATION OF LINGUISTIC KNOWLEDGE

Multiple categories of linguistic knowledge are needed for proper treatment of free text utterances.[1] Depending on the treatment and the methods which underlined it, the accent is moving from one category to others. However, the most elaborated system relies on all categories. The following categories have been recognised:

- vocabulary and its syntax,
- concepts and their typology,
- semantic co-occurrences,
- conceptual schemata and frames.

This broad categorisation is helpful in the present context, but should be further refined when considering actual implementations. Numerous other rules and linguistic information are necessary for NLP tools.

The vocabulary and the attached syntactic information is the basis of Natural Language Processing. Nothing can be done without large lexicons. But a first difficulty appears as soon as we want a multilingual lexicon: how to link together the same words in different languages? The basic answer is to link them through an abstract entity conveying the meaning of these words, what we call a concept.

At the next step, when a couple of thousand concepts have been collected, there is a need to organise them. This is what is often called a typology of concepts or an ontology. Implementations oriented towards nomenclature and classification use single inheritance hierarchy schemes. Other implementations are based on a compositional model of the domain and are in a better position to provide the semantic information needed for NLP tools.

Ongoing works on domain representation are limited. They should be able to issue from their model a set of relevant semantic rules for NLP tools. This is the case of GALEN.[2] The semantic rules are intended to control the aggregation of near words in sentences, as well as to specify what is sensible to say and what is necessary to say about a given concept. A typical example is: Fracture hasLocation Bone. Some forms of semantic rules are necessary for a semantic validation of medical texts. In UMLS the corresponding semantic rules are named syntagmatic expressions.[3]

Finally, conceptual frames are necessary to put together the different parts of a sentence. They are a kind of script of typical situations in the domain of interest. Some frame-like approaches in the medical domain exist and they have the merit of exploring and opening this road.[4]

In this paper we focus on these four categories when considering different sources. All categories are important depending on the task to perform. We will see, however, that vocabulary sources (first category) are more developed than semantic sources (categories

0195—4210/96/$5.00

2 to 4). Therefore, we will insist on the need for completing the offer of linguistic sources in the last three categories and what solutions are foreseen in the near future.

THE MULTILINGUAL ASPECT

There is no doubt that any linguistic source in the future will be a mix of syntactic and conceptual information, like two dimensions in the plan. The general reality is, however, a three dimensional reality: multilinguality is the necessary third dimension. The growth of Internet is there to give evidence of multiple languages, and certainly no more than 20 % of medical records worldwide are written in English. An ethical point of view on this problem is just another reason to think « multilingual » from the start.

Multilinguality is in fact already present even when considering a single language. It is known from several authors that a valid representation of natural language texts should be independent of the language; it should be abstracted (literally to take out of something) from that language in order to convey its meaning without disturbance from the way to express it. Such a representation has often the name of « interlingua »[4]. This kind of common representation is the natural bridge between languages and in fact represents the most difficult part of representing multiple languages. Our experience has shown that a much clearer view of the entire process is obtained when working with a minimum of two languages.[5] Working currently with 5 languages (English, French, German, Italian and Dutch) gives us more confidence in the future outcome.

When organising a lexicon, we decided to center it around concepts, and for each of them to select the annotations (the words) which represent them best in multiple languages.[6] This means a strong constraint the cost of which in terms of manpower resources is considerable: each word in the lexicon must have a corresponding concept. Knowing that a concept is not a free floating apex, but it has to be linked in some typology of the domain, we have to manage such an entity and this is what is called « the modelling process » whatever the degree of refinement is that it has been decided to achieve.

AVAILABLE SOURCES

In fact, any written text is a source from which linguistic knowledge may be extracted. Such texts are everywhere, often in machine readable form. However, for the same reason that water is essential but not sufficient for life, texts are not sufficient for linguistic knowledge acquisition as long as you don't have already a substantial part of this knowledge. The missing ingredient is the intelligence of the domain which is necessary to organize the collection of knowledge.

In this section we want to consider different corpus of linguistic information which are available and try to evaluate their richness in order to feed a Medical Linguistic Knowledge Base (MLKB)[1].

The UMLS Source

The UMLS knowledge sources provide huge amounts of linguistic information readily available to the medical community. Nevertheless, the structure and the organisation of this knowledge is somewhat a kind of compromise between different approaches.

An extensive vocabulary of some 60'000 entries is provided through the Specialist lexicon.[7] In addition, the metathesaurus gives words and expressions in accordance to the semantic network (132 semantic types), principally in English. Despite there are entries in French, Spanish and Portuguese, their usage is limited by lack of coverage and by the fact that accents are missing! The hierarchy of concepts (MeSH) covers quite a large part of the medical domain but it is not a compositional model. All members in the hierarchy are defined by ISA links from an ancestor, and not by composition of two members. This means that the genericity coming out of multiple inheritance mechanisms is not there, and this may lead to an explosion of the number of semantic rules. These are eventually called in UMLS syntagmatic expressions,[3] but have not yet been made available. Frame knowledge about prototypical situations is not part of the current UMLS release.

Our current experiments to use UMLS in order to populate an existing model like GALEN are promising. Starting from a semantic category which has already been modelled in the GALEN typology - for example BodySubstance - we have been able to extract by program 180 new concepts from the Metathesaurus, together with their English and French annotations. These new concepts appears as a flat list, but they may be later refined by composite modelling. Being part of the typology, they inherit all the properties of their ancestors and this is immediatly used by NLP tools.

In summary, UMLS qualitatively and quantitively covers the first two categories of linguistic knowledge. The last two are not really there yet. Compared to SNOMED the richness of information is higher, but the coverage is lower.

The GALEN Source

The GALEN consortium is working together since 1992 and has produced, using the GRAIL representation language, a general model of medicine with nearly 6'000 concepts. Being compositional this approach gives a substantial degree of detail for concepts and a fine grain of knowledge representation.[2] This model is compatible with conceptual graphs representation and the view it gives on a specified concept makes it analogous to a frame system. The drawback of the GALEN approach is the difficulty to grow and to reach a ready-to-use coverage of the medical domain.

The GALEN representation favors the building of multilingual lexicons around central concepts. It has already proved to be a good recipient for that. It provides a highly structured typology of concepts where linguistic rules are attached and inherited along the multiple paths hierarchy. In addition, the semantic rules are immediately derived from the GALEN sensible statements. The fourth category of linguistic knowledge is partially present in this system, but some manual processing would have to be done to extract it.

In summary, GALEN provides all four categories of knowledge and is qualitatively a good source. However, its coverage is between 1/10 to 1/4 of what is needed in clinical applications except in very narrow medical specialities. Moreover, he complexity of the modelling scheme makes the practical usage of GALEN knowledge difficult.

The MED Source

The Medical Entity Dictionary (MED) is a frame-like model of the medical domain with a specific objective of providing a controlled vocabulary for medical applications. It is based on the UMLS semantic network and has been progressively extended in order to meet the needs of ancillary clinical systems.[8]

The MED system provides a large vocabulary and a structured typology with multiple inheritance. Semantic rules for NLP are not directly available in this system. Its frame-based design makes it is a relevant source of frames of sensible situations.

The SNOMED Source

SNOMED International distributes a multilingual edition with an extensive coverage of the vocabulary and concepts of medicine. It is available on machine readable form. It is a systematic nomenclature and therefore presents the advantage of a large coverage. Its multiaxial approach makes SNOMED a structured source of knowledge which has the potential of being beneficial to NLP systems.

SNOMED is today the largest source of medical vocabulary (132'643 entries) organised in a systematic way. It is intended to be multilingual and annotations in different languages are linked by their common code. The SNOMED hierarchy does not directly allow for compositionality, but usage of SNOMED allows compositionality of terms under the form of a sequence of entries from different axes, but such a combination is under user control and responsibility and it may be perfect nonsense (i.e. M-12100 T-C2400). Semantic rules and conceptual schemata are not available. Nevertheless, some attempts have been done in this direction and have to be mentioned.[9] The D axis has the value of a frame-like model.

In summary, SNOMED gives the best extensive coverage of vocabulary in a multilingual fashion and a good contribution to the structure of concepts in a multiaxial typology, but it lacks the last two categories of linguistic knowledge.

The Textbook Source

Different textbooks are available in machine-readable form, but typically let us take the example of the Harrison of Internal Medicine which has been published as a CD-ROM. This is a huge corpus of descriptive medical texts. This is definitely an important source of linguistic knowledge as soon as we know how to extract it. We can also consider for the same purpose the CD-ROM version of the New England Journal of Medicine.

It is possible to extract all the words in such a textbook. This approach may be of interest in order to acquire a multiple language vocabulary. Concepts are certainly difficult to extract from a textbook. This is also the case for conceptual schemata.

The main interest of textbooks is certainly the extraction of semantic rules. Having this objective in mind, we are presently experimenting with a program looking for co-occurrences of neighbour words in different syntactic situations like noun-adjective or noun-noun complements. Starting, for example, from a noun and looking at all its possible adjectives, we group them according to their concepts and their proximity in the typology. Each group may be resolved in the best situation by a single semantic rule attached to a common ancestor of the concepts in the group and by a few rules in other cases. A large part of this process may be automatised and the need for man hours will be greatly reduced compared to the final benefit.

The Large Corpora Source

Large corpora are, for example, the set of all discharge summaries written and available in machine-readable form in a given hospital. They play the same role as textbooks and they are ready to provide the same type of information. In addition, such a situation is a chance to capture local oddities and specific medical jargons (which have been eliminated from a textbook).

The Classification Source

A classification like ICD-10 or ICD-9 CM is another source of linguistic knowledge because the used terms are supposed to be representative of some international medical practices. The advantage of such a nomenclature is that it is published in different languages. A carefully tuned program may be able to find corresponding words when analysing entries of the same code in different languages. This kind of grouping is an indication for a new concept. This process may be largely automatised, but it needs a parser in any new language.

RESULTS

Our current work addresses multiple knowledge sources and certainly we want the best to be extracted from each of them. Table 1 gives a summary of the past and ongoing attempts and the kind of outcomes either achieved or foreseen. Shaded areas represent

	Vocabulary	Typology	Semantic rules	Frames
UMLS	> 60'000 terms, principally English with contributions in French, Spanish and Portuguese, *ready to use*	ISA hierarchy (MesH), single inheritance, semantic network of 132 terms, *not directly applicable*	mention of syntagmatic expressions, *not yet available*	*not available*
SNOMED	> 132'000 terms, 12 languages to be released (when ?), *ready to use*	12 axes with ISA hierarchy, single inheritance, no network, *not directly applicable*	large corpus of expressions, *to be extracted*	The Disease axis (D) acts as a kind of frame-based model, *to be extracted*
ICD9-CM ICD-10	selected vocabulary, multilingual by comparison between different languages, *to be extracted*	ISA hierarchy, single inheritance, *not directly applicable*	*not available*	selected expressions may be a source of conceptual schemata, *to be extracted*
GALEN	> 4'000 terms in 5 languages: English, French, German, Italian, Dutch *ready to use*	compositional model, multiple inheritance, modelling tools, *ready to use*	1'300 rules, directly available from model, *ready to use*	not directly available, but deducible from the model, *to be extracted*
MED	> 32'000 terms in English, *ready to use*	multiple inheritance hierarchy, based on UMLS semantic network, *ready to use*	*not available*	Frame-oriented approach providing numerous frames, *ready to use*
Text Books	unlimited number of words, each source is monolingual, *to be extracted*	*not available*	an unlimited number of rules are implicite and may be sorted on common concepts, *in progress*	selected text books may be a source of conceptual schemata, *to be extracted*
Corpora	selected vocabulary, each source is monolingual, *to be extracted*	*not available*	an unlimited number of rules are implicite and may be sorted on common concepts, *in progress*	selected text may be a source of conceptual schemata, *to be extracted*

Table 1: Comparison of different knowledge sources in respect to categories of linguistic knowledge

the most relevant sources to our point of view. This appreciation is given regarding the strict NLP point of view knowing that these sources have been developed with other objectives in mind.

A few comments result from this table. The double line separation shows the specific roles of the knowledge sources. « Extensive sources » like UMLS and SNOMED gives the vocabulary and ISA typology. « Intensive sources » like GALEN, MED or medical texts in general contribute to the necessary model of medicine we need for NLP and other applications. In order to properly work on the extraction from intensive sources, preparation runs are necessary on the extensive sources. Without some coverage of the domain, there is no hope to extract qualitative knowledge from large corpora.

In order to illustrate the dependance of NLP tools on semantic knowledge as issued from a model of the domain, we can mention the example of two significant systems: one is working on clinical radiology reports[10] and relies on MED; the other is working on discharge summaries[11] and relies on GALEN. In both systems the NLP functionalities are grounded on the knowledge issued from the underlying model: the more developed the model, the more advanced the NLP capabilities.

INTERNET USAGE

Linguistic knowledge in a multilingual context is in perpetual evolution. No single group may assume the responsibility of the entire process of corrective maintenance, completion of the knowledge, evolutive maintenance and improvement of the tools and methods. The only solution is cooperative development and maintenance.

The GALEN-IN-USE consortium is developing and experimenting now with a system of cooperative modelling as well as cooperative linguistic acquisition. This means that members of the European Federation of Coding Centers (EFCC) will contribute to the building of new knowledge through Internet.

CONCLUSION

Undoubtly, all mentioned knowledge sources (and others) are useful for NLP and the major lesson from this round trip is that they are complementary. When extracting knowledge from multiple sources one is faced to the problem of multiple if not incompatible representations. One way to supposedly solve this problem is to add just another representation at the risk of augmenting the confusion for future users. Nevertheless, we have learnt that we are far from a standard for concepts representation in medicine.

UMLS and SNOMED are the most quantitively valuable sources due to their impressive size. They are in fact well known and they will have convenient resources in the years to come. But may be they have reached a significant size because they concentrate more resources on the vocabulary and taxonomy aspects than to the modelling aspect. NLP in the future needs more qualitative knowledge as issued from a model of the domain. Any future effort in this direction, like MED or GALEN, may soon trigger relevant results and new clinical applications.

Acknowledgments

The present work has been realised in the GALEN-IN-USE consortium.

References

1. Baud RH, Lovis C, Rassinoux AM, Michel PA, Alpay L, Wagner JC, Juge C, Scherrer JR, Towards a Medical Linguistic Knowledge Base. Greenes RA et al. (editors) MEDINFO'95 proceedings, 1995, North Holland, pp 13-17.

2. Rector AL, Coordinating Taxonomies: Key To Re-usable Concept Representations. In proceedings of Artificial Intelligence in Medicine Europe, AIME'95, Barahona P & al (editors) Spinger Verlag, 1995, pp 17-28.

3. McCray AT, Nelson SJ. The Representation of Meaning in the UMLS, Meth. Inform. Med., 1995; 34: 193-201.

4. Masarie FE, Miller RA, Bouhaddou O, Giuse NB, Warner HR. An Interlingua for Electronic Interchange of Medical Information: Using frames to Map between Clinical Vocabularies. Computers and Biomedical Research 24, 379-400 (1991).

5. Wagner JC, Solomon WD, Michel PA, Juge C, Baud RH, Rector AL, Scherrer JR. Multilingual Natural Language Generation as Part of a Medical Terminology Server. Greenes RA et al. (editors) MEDINFO'95 proceedings, 1995, North Holland, pp 100-4.

6. Baud RH, Rassinoux AM, Wagner JC, Lovis C, Juge C, Alpay LL, Michel PA, Degoulet P, Scherrer JR. Representing Clinical Narratives Using Conceptual Graphs. Meth. Inform. Med., 1995; 34: 176-86.

7. McCray AT, Srinivasan S, Browne AC. Lexical Methods for Managing Variation in Biomedical Terminologies. 18th Annual Symposium on Computer Applications in Medical Care, 1994, Washington, pp 235-9.

8. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB, Knowledge-based Approaches to the Maintenance of a large controlled Medical Terminology, J Am Med Informatics Assoc, 1994; 1(1): 35-50.

9. Campbell KE, Musen MA. Representation of clinical Data using SNOMED III and conceptual Graphs. 16th Annual Symposium on Computer Applications in Medical Care, 1992, Washington, pp 354-358.

10. Friedman C, Cimino JJ, Johnson SB. A Schema for Representing Medical Language Applied to Clinical Radiology. J Am Med Informatics Assoc, 1994;1(3):233-48.

11. Rassinoux AM, Wagner JC, Lovis C, Baud RH, Rector AL, Scherrer JR. Analysis of Medical Texts Based on a Sound Medical Model. 19th Annual Symposium on Computer Applications in Medical Care, 1995, New Orleans, pp 27-31.